Research Report

Fitting the Structured General Diagnostic Model to NAEP Data

Xueli Xu

Matthias von Davier

May 2008

ETS RR-08-27

ETS

Listening. Learning. Leading.®

Fitting the Structured General Diagnostic Model to NAEP Data

Xueli Xu and Matthias von Davier
ETS, Princeton, NJ

May 2008

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Copyright © 2008 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, LISTENING. LEARNING. LEADING.,
and TOEFL are registered trademarks of Educational Testing
Service (ETS).

Abstract

Xu and von Davier (2006) demonstrated the feasibility of using the general diagnostic model (GDM) to analyze National Assessment of Educational Progress (NAEP) proficiency data. Their work showed that the GDM analysis not only led to conclusions for gender and race groups similar to those published in the NAEP Report Card, but also allowed flexibility in estimating multidimensional skills simultaneously. However, Xu and von Davier noticed that estimating the latent skill distributions will be much more challenging with this model when there is a large number of subgroups to estimate. To make the GDM more applicable to NAEP data analysis, which requires a fairly large subgroups analysis, this study developed a log-linear model to reduce the number of parameters in the latent skill distribution without sacrificing the accuracy of inferences. This paper describes such a model and applies the model in the analysis of NAEP reading assessments for 2003 and 2005. The comparisons between using this model and the unstructured model were made through the use of various results, such as the differences between item parameter estimates and the differences between estimated latent class distributions. The results in general show that using the log-linear model is efficient.

Key words: General diagnostic model (GDM), group characteristics, item responses, large survey data, log-linear modeling

1. Introduction

Located latent class (LLC) models are alternatives to item response theory (IRT) models (Lord & Novick, 1968) in analyzing item responses. For LLC models, the latent abilities are conceptualized as an ordered or unordered set of a finite number of fixed classes (Haberman, 1979; Lazarsfeld & Henry, 1968). In IRT, the latent abilities are assumed to be continuous in a countable set. Though these two types of modeling make different assumptions about the latent abilities, they have something in common. For example, the marginal maximum likelihood estimates (MMLEs) of the item parameters will be identical for these two modeling frameworks when the location and number of the ordered latent classes are chosen to reflect a sufficient number of appropriately spaced points on the latent ability distribution (Formann, 1992; Heinen, 1996; von Davier & Yamamoto, 2004). Some arguments have been advanced that LLC models are more efficient than IRT models. For example, Laird (1978) showed that the estimated latent distribution obtained from MMLEs is a discrete distribution even if the true distribution is continuous. In addition, Follmann (1988) and Haberman (2005) showed that, at most, $(T + 2)/2$ discrete latent classes can be identified in a test consisting of total T items when using the Rasch model. The number of identifiable latent classes might be even smaller in other models, such as the two-parameter logistic (2PL) or the three-parameter logistic (3PL) models (Haberman, 2005). These arguments imply that LLC models are useful and meaningful alternatives to IRT models. In addition, LLC models may be more efficient than IRT models in cases where small numbers of items are administered, and/or in situations where items measuring multidimensional latent abilities are administered.

An immediate application of LLC models is cognitive diagnosis where the inferences about individuals are made in the form of multiple discrete latent abilities. Examples of such approaches include the general diagnostic model (GDM; von Davier, 2005), the fusion model (FM; Hartz, 2002), and the rule-space methodology (RSM; Tasuoka, 1995), as well as latent response models and multiple-classification latent class models (Junker & Sijtsma, 2001; Maris, 1995, 1999). However, one potential problem exists for such models if no constraints are imposed on the structure of the latent skill space. It is not difficult to observe that the skill space composed by the combination of the discrete latent skills can be easily expanded as the number of skills as well as the levels for each skill increase. For example, suppose a verbal test is designed to measure 10 skills. If no constraints are put into the structure of the latent skill space, this amounts to

estimating $2^{10} - 1$ (i.e., 1,023) probabilities of the latent ability profiles. The same number of 10 dimensions with three levels each amounts to $3^{10} - 1$ (i.e., 59,048) parameters to estimate. The vast majority of data sets will be sparse relative to this large number of latent classes. This observation was motivation use a parametric form to reflect main features of the latent skill space.

It seems natural to use log-linear modeling for the latent class distribution since the latent variables in LLC models are discrete. This study uses real data examples to demonstrate the log-linear model for the latent class distribution will not affect efficiency in estimating item parameters and marginal skill distributions. This paper is organized as follows: Section 2 introduces the log-linear model, the GDM, their related notation, and their likelihood. Section 3 describes the EM algorithm used to estimate the model parameters. The real data analyses are explored in section 4, followed by a brief discussion in section 5.

2. Structured General Diagnostic Model

The GDM coupled with a log-linear model for the latent skill space is used in this study. This model is referred to as the *structured GDM*. The sections below introduce the necessary notations and guiding assumptions for the structured GDM. In contrast to approaches that assume a lower dimensional structure underlies the multidimensional skill structure (de la Torre & Douglas, 2004), the proposed model smooths the latent-class distribution by applying a log-linear model. The proposed model does not reduce dimensionality and thus stays with the number of dimensions assumed when the test was designed. However, lower dimensional models could be specified directly in the GDM framework if one wanted to consider them. Then, the fit of two or more alternative models, assuming lower and higher dimensional structures, can be compared directly.

2.1 Notation- and Model-Based Likelihood

Let N denote the total number of examinees in the sample, and M the total number of items. Let h be the index for the latent classes composed of combinations of latent skills and H the total number of latent classes. Let S_j be the number of response categories for item j, and $s(ij)$ be the score of person i in responding to item j.

Suppose $\underline{x}_i = (x_{i1}, \ldots, x_{iM})$ is the response vector of examinee i. Let $P(\underline{x}_i|h)$ denote the probability of the response pattern given the latent class h, and $P(h)$ the probability of the latent class h, for $h = 1, \ldots, H$.

Then, the marginal likelihood is given by:

$$L = \prod_{i=1}^{N} [\sum_{h=1}^{H} (P(\underline{x}_i|h)P(h))].$$

If each person's latent class [i.e., person i is from class h, denoted as $h(i)$] is known, the likelihood can be written as

$$\begin{aligned}
L &= \prod_{i=1}^{N}\prod_{h=1}^{H}[P(\underline{x}_i|h)P(h)]^{I(h(i)=h)} \\
&= \prod_{i=1}^{N}\prod_{h=1}^{H}[\prod_{j=1}^{M} P(x_{ij}|h)P(h)]^{I(h(i)=h)} \\
&= \prod_{i=1}^{N}\prod_{h=1}^{H}[P(h)\prod_{j=1}^{M}\prod_{s=1}^{S_j} P(x_{ij}=s|h)^{I(s(ij)=s)}]^{I(h(i)=h)} \\
&= \prod_{i=1}^{N}\prod_{h=1}^{H} P(h)^{I(h(i)=h)} \prod_{i=1}^{N}\prod_{h=1}^{H}\prod_{j=1}^{M}\prod_{s=1}^{S_j} P(x_{ij}=s|h)^{I(s(ij)=s)I(h(i)=h)} \\
&= \prod_{h=1}^{H} P(h)^{n(h)} \prod_{h=1}^{H}\prod_{j=1}^{M}\prod_{s=1}^{S_j} P(x_j=s|h)^{n(j,h,s).} \qquad (1)
\end{aligned}$$

Note that $n(h)$ stands for the numbers of students who are in latent class h, and $n(j,h,s)$ represents the number of students who are in latent class h, and score in category s on item j. The log-likelihood follows by taking the log of equation (1):

$$l = \log L = \sum_{h=1}^{H} n(h)\log P(h) + \sum_{h=1}^{H}\sum_{j=1}^{M}\sum_{s=1}^{S_j} n(j,h,s)\log P(x_j=s|h). \qquad (2)$$

Notice that the first term in this log-likelihood contains only the information about population parameters, while the second term contains information about item parameters.

The parametric models used in this study for $P(h)$ and $P(x_j = s|h)$ are introduced in the next two sections.

2.2 A Log-Linear Model for Multidimensional Attribute Spaces

Suppose that a total number of K attributes are measured by one test. Then the latent classes are generated by all possible combinations of these K attributes. In connection with previous notation, h is an index of latent classes realized by $(\theta_1, \theta_2, \ldots, \theta_K)$. Here, θ_k, for $k = 1, \ldots, K$, is a discrete random variable with m real valued levels for each skill. Let $n(\theta_1, \ldots, \theta_K)$ denote the

joint count of these K discrete random variables. A multidimensional log-linear model used in describing the latent skill space is given by:

$$\log n(\theta_1, \ldots, \theta_K) = \beta_{(0)} + \sum_{k=1}^{K} \sum_{l=1}^{max(l)} \beta_{(1)kl}\theta_k^l + \sum_{c=1}^{K-1} \sum_{k=c+1}^{K} \beta_{(2)kc}\theta_c\theta_k, \tag{3}$$

where l is the index for the number of moments. The maximum value of l could be 1, 2 and 3, depending on the levels specified for the latent skills. For example, if a skill only has two levels, $max(l) = 1$. If a skill has three levels, $max(l) = 2$. Otherwise, $max(l) = 3$. Including up to a third of latent variables enables one to capture skewness of the latent class distribution.

If $n(\theta_1, \theta_2, \ldots, \theta_K)$ were observable, either the Newton-Ralphson (N-R) method or the iterative weighted least square (IWLS) method could be used to estimate the parameters $\underline{\beta_{(0)}}$, $\underline{\beta_{(1)}}$, and $\underline{\beta_{(2)}}$. Details on these two methods can be found in the book by Agresti (2002). These two methods will lead to the same estimates of the parameters as well as the same estimation errors (Birch, 1963; Lang, 1996; Palmgren, 1981). However, since $n(\theta_1, \theta_2, \ldots, \theta_K)$ are not observable, the parameters have to be calculated iteratively in an algorithm suitable for incomplete data, for example the EM algorithm.

2.3 Brief Introduction to the GDM

The GDM (von Davier, 2005) has been developed and applied in specific analyses of several large scale assessment data sets, such as the National Assessment of Educational Progress (NAEP) data profile scoring (Xu & von Davier, 2006) and TOEFL® iBT data analysis (von Davier, 2005). A compensatory GDM suitable for dichotomous and polytomous ordinal items is given by:

$$\log P(x_j = s|\theta_1, \theta_2, \ldots, \theta_K) = \alpha_j(\beta_{js}, \gamma_{jk}) + \beta_{js} + \sum_{i}^{K} s\gamma_{jk}\theta_k q_{jk}, \tag{4}$$

where α_j is a normalizing term. This model will be revisited in the section on estimation.

3. Estimation: EM Algorithm

An EM algorithm suitable for estimating the parameters in the structured GDM is given below. In the E step, preliminary parameter estimates are used to calculate the expected counts, which serve as the basis for the required statistics in the M step. Then the M step iteratively improves on parameter estimates for the structured GDM using standard maximization methods. The E step and the M step are executed alternatively until some convergence criterion is reached.

3.1 E Step

With initial values of both item parameters and population parameters, the posterior distribution $P(h|\underline{x}_i) = P(\theta_{1(h)}, \theta_{2(h)}, \ldots, \theta_{K(h)}|\underline{x}_i)$ can be easily calculated according to Bayes' theorem. The two required types of expected counts are then obtained through the following two equations:

$$\hat{n}(h) = n(\theta_{1(h)}, \theta_{2(h)}, \ldots, \theta_{K(h)}) = \sum_{i=1}^{N} P(\theta_{1(h)}, \theta_{2(h)}, \ldots, \theta_{K(h)}|\underline{x}_i) \quad \text{and}$$

$$\hat{n}(j, h, s) = \sum_{i=1}^{N} I(x_{ij} = s) P(\theta_{1(h)}, \theta_{2(h)}, \ldots, \theta_{K(h)}|\underline{x}_i).$$

3.2 M Step

Two sets of parameters will be updated in the M step. Therefore, two stages are designed by divide-and-conquer technique.

3.3 M Step, Stage 1

Update the parameters in the log-linear model.

Let $\mathbf{Z}$ be a design matrix that has the following form when there are more than three levels for each skill:

$$\mathbf{Z}_{H\times G} = \begin{pmatrix} \mathbf{1} & \mathbf{A} & \mathbf{B} \end{pmatrix},$$

where $\mathbf{1}$ is a vector with all elements being 1. $\mathbf{A}$ is a matrix of dimension $H \times 3K$, with elements being the main effects of the K discrete random variables:

$$\mathbf{A} = \begin{pmatrix} \theta_{1(1)} & \ldots & \theta_{K(1)} & \theta^2_{1(1)} & \ldots & \theta^2_{K(1)} & \theta^3_{1(1)} & \ldots & \theta^3_{K(1)} \\ \theta_{1(2)} & \ldots & \theta_{K(2)} & \theta^2_{1(2)} & \ldots & \theta^2_{K(2)} & \theta^3_{1(2)} & \ldots & \theta^3_{K(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta_{1(H)} & \ldots & \theta_{K(H)} & \theta^2_{1(H)} & \ldots & \theta^2_{K(H)} & \theta^3_{1(H)} & \ldots & \theta^3_{K(H)} \end{pmatrix}.$$

$\mathbf{B}$ is a matrix of dimension $H \times \frac{K(K-1)}{2}$, with elements being the first-order correlation of these variables.

$$\mathbf{B} = \begin{pmatrix} \theta_{1(1)}\theta_{2(1)} & \ldots & \theta_{1(1)}\theta_{K(1)} & \theta_{2(1)}\theta_{3(1)} & \ldots & \theta_{K-1(1)}\theta_{K(1)} \\ \theta_{1(2)}\theta_{2(2)} & \ldots & \theta_{1(2)}\theta_{K(2)} & \theta_{2(2)}\theta_{3(2)} & \ldots & \theta_{K-1(2)}\theta_{K(2)} \\ \vdots & \vdots & \vdots & \vdots & & \\ \theta_{1(H)}\theta_{2(H)} & \ldots & \theta_{1(H)}\theta_{K(H)} & \theta_{2(H)}\theta_{3(H)} & \ldots & \theta_{K-1(H)}\theta_{K(H)} \end{pmatrix}.$$

Therefore, the design matrix $\mathbf{Z}$ is an $H \times G$ matrix, with H being the total number of attribute patterns and G equal to $1 + \sum_{k=1}^{K} Min(3, nlevel(k) - 1) + \frac{K(K-1)}{2}$ when the levels for each skill k are denoted by $nlevel(k)$. In this design matrix, $\theta_{k(h)}$ represents the value of θ_k in the attribute pattern h. For example, in an attribute pattern $h = \{1011\}$, $\theta_{2(h)} = 0$ and $\theta_{3(h)} = 1$.

Let $\underline{\beta} = (\underline{\beta_{(0)}}, \underline{\beta_{(1)}}, \underline{\beta_{(2)}})$ be the vector of parameters that correspond to this design matrix.

The log-linear model in (3) can be written in terms of the design matrix and the population parameter vector:

$$\log n(\theta_1, \ldots, \theta_K) = \mathbf{Z}\beta.$$

The iterative weighted least square (IWLS) estimator of $\underline{\beta}$ is given by:

$$\underline{\hat{\beta}}_{IWLS} = (\mathbf{Z}^T\mathbf{V}\mathbf{Z})\mathbf{Z}\mathbf{V}\log(\underline{\hat{n}(h)}),$$

where $\mathbf{V} = diag\{\hat{n}(h)\}$ and $\underline{\hat{n}(h)}$ is vector with elements $\hat{n}(h)$. To make this vector estimable, three conditions have to be met: (a) $K \geq 1$, (b) $H \geq 3$, and (c) $H > G$. The estimation error matrix of $\underline{\hat{\beta}}_{IWLS}$ is:

$$cov(\underline{\hat{\beta}}_{IWLS}) = [\mathbf{Z}^T V \mathbf{Z}]^{-1}.$$

3.4 M Step, Stage 2

The second phase of the M step updates the parameters in the GDM in (4). This probability function belongs to the exponential family, with parameters β_{js} and γ_{jk}s, for $s = 1, \ldots, S_j$ and $k = 1, \ldots, K$. For parameter β_{js}, the sufficient statistic is $I(x_j = s)$. For parameter γ_{jk}, the sufficient statistic is $s\theta_k q_{jk}$. It then follows that:

$$\frac{\partial l}{\partial \beta_{js}} = \sum_{h=1}^{H} \hat{n}(j, h, s)(1 - P(x_j = s|\theta_{1(h)}, \theta_{2(h)}, \ldots, \theta_{K(h)}))$$

and

$$\frac{\partial l^2}{\partial \beta_{js}^2} = \sum_{h=1}^{H} \hat{n}(j, h, s)P(x_j = s|\theta_{1(h)}, \theta_{2(h)}, \ldots, \theta_{K(h)})(1 - P(x_j = s|\theta_{1(h)}, \theta_{2(h)}, \ldots, \theta_{K(h)})).$$

The same logic is followed to calculate the first and second derivative of γ_{jk}s:

$$\frac{\partial l}{\partial \gamma_{jk}} = \sum_{h=1}^{H} \sum_{s=1}^{S} \hat{n}(j, h, s)\theta_{k(h)} q_{jk}[s - P(x_j = s|\theta_{1(h)}, \theta_{2(h)}, \ldots, \theta_{K(h)})],$$

and

$$\frac{\partial l^2}{\partial \gamma_{jk}^2} = \sum_{h=1}^{H} \sum_{s=1}^{S} n(j,h,s)\theta_{k(h)}^2 q_{jk}^2 [E(s^2) - (E(s))^2].$$

The N-R algorithm is then used to obtain updated estimates:

$$\gamma_{jk}^{(n+1)} = \gamma_{jk}^{(n)} - (\frac{\partial l^2}{\partial \gamma_{jk}^2})^{-1} \frac{\partial l}{\partial \gamma_{jk}}$$

and

$$\beta_{js}^{(n+1)} = \beta_{js}^{(n)} - (\frac{\partial l^2}{\partial \beta_{js}^2})^{-1} \frac{\partial l}{\partial \beta_{js}}.$$

The EM algorithm is stopped when certain convergence criteria are met. The convergence criteria used are typically based on the convergence of parameter estimates as well as the log-likelihood. In the *mdltm* software (von Davier, 2005) used for the analyses presented here, both criteria have to be met before the algorithm stops the maximization process.

4. NAEP Data Analysis

In this study, real data examples are used to demonstrate the performance of the structured GDM, compared to the unstructured GDM. Recall the definition of the structured GDM in the beginning of section 2. The difference between the structured and unstructured GDM lies in the use of a log-linear model for the latent skill space, which is used in the structured GDM but not in the unstructured GDM. This implies that the unstructured GDM has more skill space parameters to estimate than the structured GDM. Given sufficient data, it is obvious that the unstructured GDM will lead to a larger likelihood than the structured GDM because it uses more parameters. However, if the log-linear model of the latent skill space captures the main features of the distribution, the use of the structured GDM will not lose efficiency in estimating the latent class distributions and the item parameters. In addition, the structured GDM might result in better model fits in terms of improved information indices, since it is more parsimonious.

4.1 Analysis Plan

Two reading data sets from the NAEP were analyzed. One data set was from the Grade 4 assessment in 2003, the other one is from the Grade 8 assessment in 2005. Two subscales with 102 items were administered to 191,271 students in Grade 4 in 2003. Three subscales with 142 items were administered to 159,449 students in Grade 8 in 2005. Each assessment included a certain proportion of polytomous items. In the analysis, the subscales defined in the frameworks were

taken as cognitive skills or attributes in the GDM framework (see Xu & von Davier, 2006, for a discussion).

The NAEP data analysis is conducted under both the single-group and the multiple-group assumptions. In a single-group analysis, only one latent skill distribution is assumed for the latent skill space; while in a multiple-group analysis, possibly different skill distributions are assumed for different subgroups represented by gender groups or racial groups.

4.2 Results: Comparing the Structured and the Unstructured GDM

The comparison between the structured and unstructured GDM is made in terms of the item parameter estimates, model fit, and marginal skill distribution estimates.

Item parameter estimates. Figure 1 presents the boxplots of the difference in item parameter estimates between the structured and unstructured GDM. Assumption 1 stands for the analysis under the single-group assumption, while Assumption 2 represents the analysis under the multiple-group model, using either gender or racial ethnicity as the grouping variable. The lighter boxplots are the results for the NAEP Grade 4 reading data of 2003, while the darker boxplots are the results for the Grade 8 reading data of 2005. It is observed that, no matter which assumption is used and which data are used, the difference between the structured and unstructured GDM is small in terms of the item parameter estimates.

Model fit. The number of parameters, the log-likelihood and the Akaike information criterion (AIC; Akaike, 1974) index are shown in Tables 1 and 4. In particular, Tables 1 and 2 are the results under the single-group assumption, while Tables 5 and 4 are under the multiple-group assumption. Two types of comparisons can be made by examining these tables: the comparison between the structured and the unstructured GDM, and the comparison between different models (i.e., two-parameter logistic/general partial-credit model [2PL/GPCM], three-level GDM, and four-level GDM). The three-level GDM stands for the case where each skill is assigned three real-valued levels, while four real-valued levels are assigned to each skill in the four-level GDM.

First look at the difference between the structured and unstructured GDM. Tables 1–4 show that these two models are similar to each other in terms of the log-likelihood and the AIC index. The reduction of parameter space in the structured GDM did not pose any difficulties in fitting the data.

Next, when the models are compared in the single-group assumption, it can be observed that

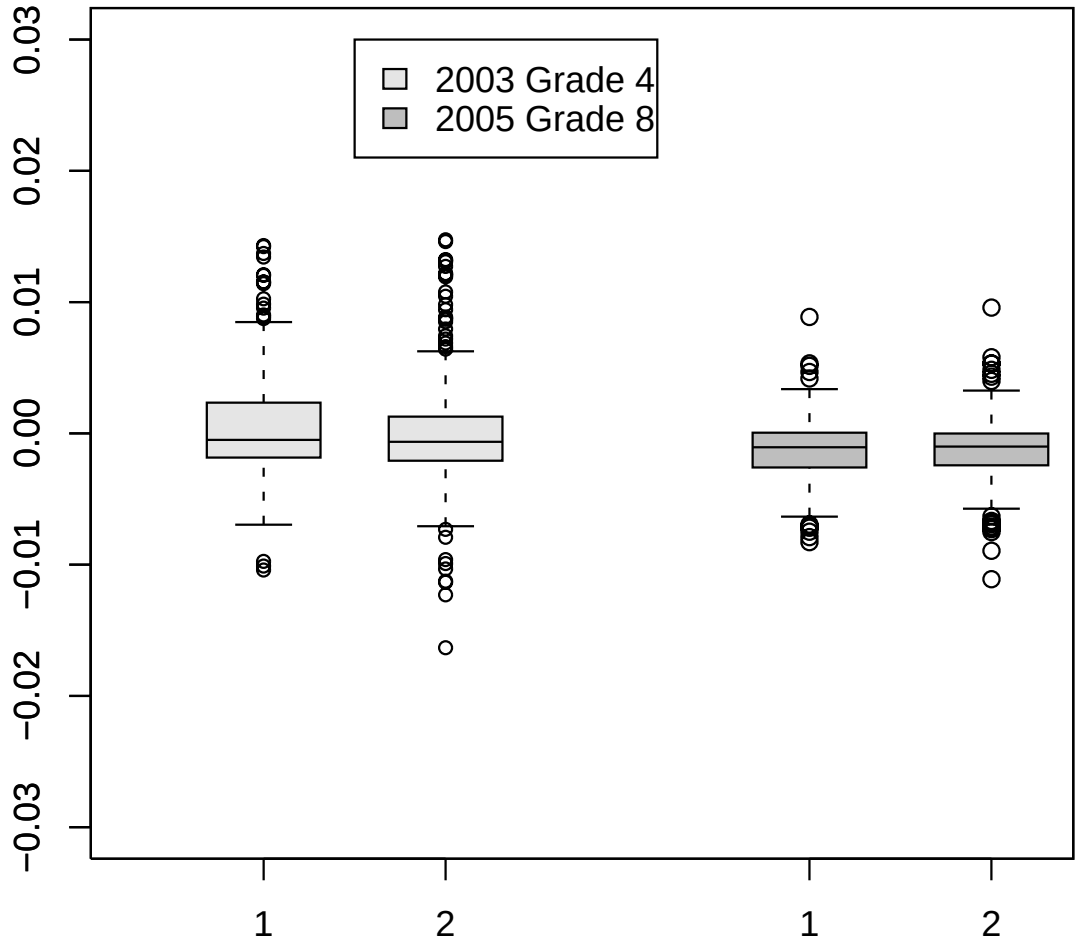

***Figure* 1. The difference between the structured and unstructured GDM.**

the four-level GDM gives the best fit under both the structured and unstructured GDM. Hence, the four-level GDM is used in multiple-group analysis in this paper.

Marginal skill distribution estimates. This type of comparison can be demonstrated using the results of the Grade 8 reading data analysis. The marginal skill distribution estimates are shown in Table 5 for the four major ethnicity groups. The results provided here are based on the analysis under the multiple-group model with ethnicity as the grouping variable. The models used in the analysis include the four-level-three-skill structured and unstructured GDM. Compared to

Table 1.
2003 ***Grade 4 Reading Data Analysis: Under Single-Group Assumption***

Model	# of parameters in unstructured GDM	# of parameters in structured GDM	Log-likelihood in unstructured GDM	Log-likelihood in structured GDM	AIC in unstructured GDM	AIC in structured GDM
2PL/GPCM	261	235	-2,124,703.89	-2,124,959.79	4,249,930	4,250,390
3-level GDM	241	240	-2,127,534.21	-2,127,539.00	4,255,550	4,255,558
4-level GDM	246	240	-2,123,603.02	-2,123,709.38	4,247,698	4,247,899

Table 2.
2005 Grade 8 Reading Data Analysis: Under Single-Group Assumption

Model	# of parameters in unstructured GDM	# of parameters in structured GDM	Log-likelihood in unstructured GDM	Log-likelihood in structured GDM	AIC in unstructured GDM	AIC in structured GDM
2PL/GPCM	359	334	-1,867,520.94	-1,867,667.05	3,735,760	3,736,002
3-level GDM	357	342	-1,869,103.90	-1,869,119.49	3,738,922	3,738,923
4-level GDM	391	342	-1,866,312.36	-1,866,356.62	3,733,407	3,733,397

Table 3.
2003 Grade 4 Reading Data Analysis: Under Multiple-Group Assumption

Model	# of parameters in unstructured GDM	# of parameters in structured GDM	Log-likelihood in unstructured GDM	Log-likelihood in structured GDM	AIC in unstructured GDM	AIC in structured GDM
4-level GDM with gender	260	247	-2,122,764.47	-2,122,875.36	4,246,049	4,246,245
4-level GDM with race	288	261	-2,112,494.57	-2,112,606.41	4,225,565	4,225,735

using the unstructured GDM, using the structured GDM leads to similar estimates for marginal probability, except for several cases highlighted in bold in Table 5. Even for these highlighted cases, the difference of 0.01 between the structured and unstructured GDM is acceptable and negligible.

Table 4.
2005 Grade 8 Reading Data Analysis: Under Multiple-Group Assumption

Model	# of parameters in unstructured GDM	# of parameters in structured GDM	Log-likelihood in unstructured GDM	Log-likelihood in structured GDM	AIC in unstructured GDM	AIC in structured GDM
4-level GDM with gender	452	353	-1,865,005.74	-1,865,062.97	3,730,915	3,730,832
4-level GDM with race	574	375	-1,858,682.43	-1,858,767.66	3,718,513	3,718,285

Table 5.
2005 Grade 8 Reading Data Analysis: The Marginal Distribution Under Multiple-Group Assumption

		Unstructured GDM				Structured GDM			
		Level 1	Level 2	Level 3	Level 4	Level 1	Level 2	Level 3	Level 4
White	Skill 1	0.06	0.19	0.38	0.37	0.06	0.19	0.38	0.37
	Skill 2	0.02	0.11	0.38	0.49	0.02	0.11	0.38	0.49
	Skill 3	0.01	0.09	0.38	0.52	0.01	0.09	0.38	0.52
Black	Skill 1	0.19	0.39	***0.31***	***0.11***	0.19	0.39	***0.32***	***0.10***
	Skill 2	0.05	0.29	0.48	0.17	0.05	0.29	0.48	0.17
	Skill 3	0.03	0.26	0.52	0.19	0.03	0.26	0.52	0.19
Hispanic	Skill 1	***0.18***	***0.33***	0.35	0.14	***0.17***	***0.34***	0.35	0.14
	Skill 2	0.05	0.27	0.48	0.20	0.05	0.27	0.48	0.20
	Skill 3	0.04	0.27	0.47	0.22	0.04	0.27	0.47	0.22
Asian-American	Skill 1	***0.07***	***0.20***	0.34	0.39	***0.06***	***0.21***	0.34	0.39
	Skill 2	0.02	0.12	0.34	0.52	0.02	0.12	0.34	0.52
	Skill 3	0.01	0.10	0.34	0.54	0.01	0.10	0.34	0.54

4.3 Results: Comparing the Single-Group and Multiple-Group Assumptions

Given the results presented above, the efficiency of using the structured GDM to fit the data has been demonstrated. That being the case, the comparisons between the single-group and multiple-group assumptions were made using the structured GDM. Under the single-group assumption, the marginal skill distribution of a subgroup was obtained through a summation of the posterior distribution of each student in that subgroup. Under the multiple-group assumption, the skill marginal distribution was estimated directly in the software program *mdltm* (von Davier, 2005). The comparisons between the single-group and multiple-group assumptions are presented in terms of item parameter and marginal skill distribution estimates.

Item parameter estimates. Figure 2 illustrates the difference in item parameter estimates under the single-group and multiple-group assumptions. The lighter boxplots represent the differences in analyzing the 2003 Grade 4 reading data under these two assumptions. The analysis under the multiple-group assumption includes the gender-group analysis and the race-group analysis labeled on the x-axis. The darker boxplots are the results from 2005 Grade 8 reading data analysis. All analysis in this figure were conducted by using the structured GDM with four levels specified for each skill. The group labeled *gender* on the x-axis represents the difference in item parameter estimates obtained by using single-group assumption and the gender-group analysis. Likewise, the group labeled *race* on the x-axis stands for the difference obtained by using the single-group and the race-group assumptions. It can be observed that the analysis under the gender-group assumption leads to item parameter estimates that are similar to those under the single-group assumption. However, the difference is larger when comparing the item parameter estimates from the single-group and the race-group assumptions.

Marginal skill distribution estimates. This section focuses on the comparison between the single-group and the multiple-group assumptions in terms of the marginal skill distribution estimates. This comparison is illustrated using the Grade 8 reading analysis. In particular, Figure 3 compares the single-group analysis to the gender-group analysis, while Figure 4 shows the differences between the single-group and the race-group analysis. Since Skills 2 and 3 have patterns that are similar to the pattern found for Skill 1, only the analysis for Skill 1 is shown in Figure 4 for illustration.

In each graph, the solid line stands for the results from the single-group analysis, while the dashed line represents the results from either the gender-group analysis or the race-group analysis. The differences are small when the single-group and gender-group analysis are compared, as shown in Figure 3. However, Figure 4 tells us a different story. This figure plots the marginal distribution estimates for four major racial groups: White, Black, Hispanic and Asian-American students. For the White and Asian-American students, the marginal distribution estimates are not affected by the choice of the single-group and race-group assumptions. However, a slight difference can be noted for the Hispanic and Black students when different analysis assumptions are chosen. Further investigation is needed to determine whether these differences are significant.

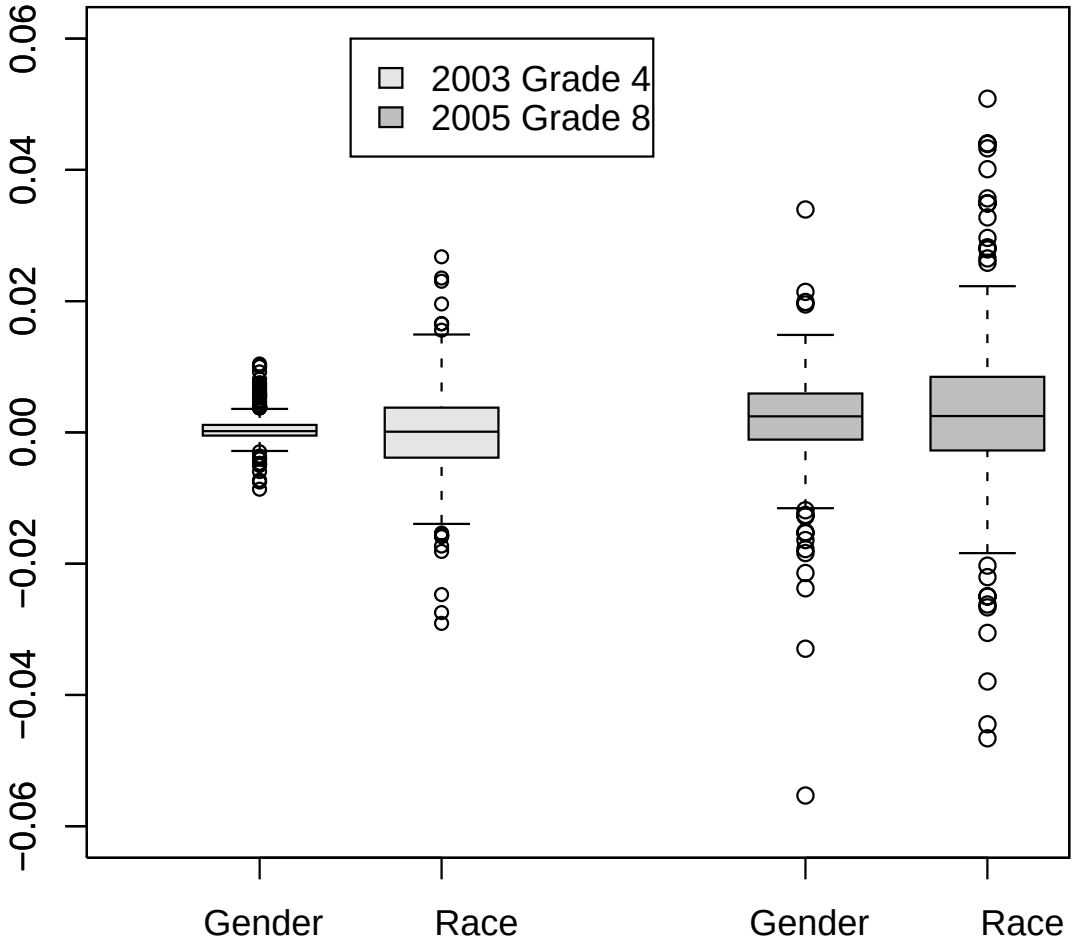

***Figure* 2. The difference between single- and multiple-group assumptions.**

5. Discussion

5.1 Structured Versus Unstructured GDM

The difference between the structured and the unstructured GDM lies in the structure employed in describing the latent skill distribution. In the structured GDM, a log-linear model was utilized to smooth the distribution of multivariate discrete latent random variables. The purpose of utilizing such a model is to improve on estimation efficiency, in addition to reducing the parameter space. The results in Figure 1 and Tables 1–4 provide the evidence that the log-linear

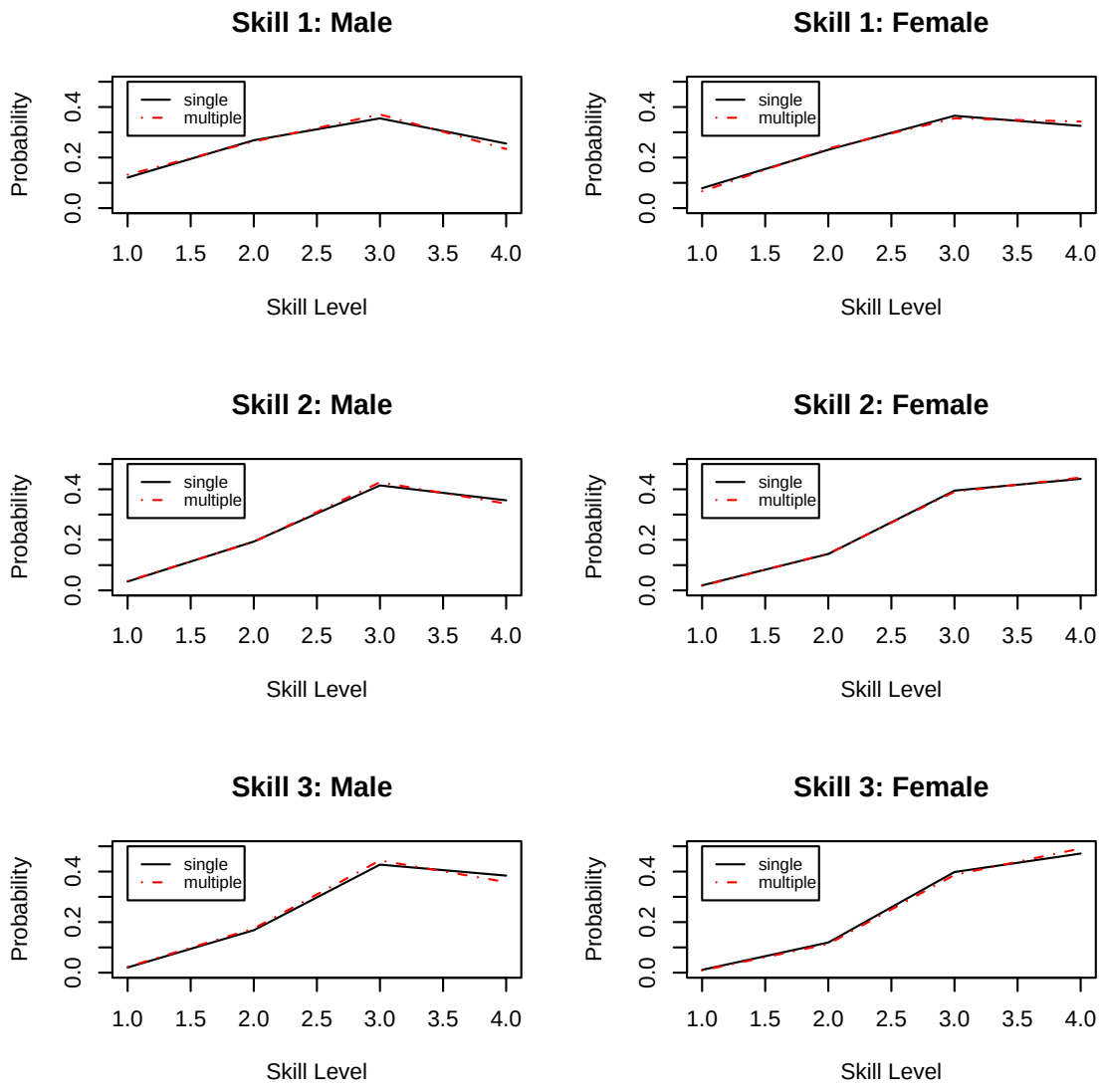

***Figure* 3. Marginal skill distributions: The single-group and gender-group analysis of the NAEP 2005 Grade 8 reading data.**

approach is capable of modeling the skill space accurately. It should be noted, however, that the log-linear models may break down when there are a fair amount of structural zero cells or cells with zero counts. For a test with highly correlated subscales, the parameter estimates might fluctuate slightly around the true value, so that the algorithm would appear not to converge. Further research needs to be conducted along this line.

In fact, the log-linear modeling of the latent attribute space is not limited to the GDM

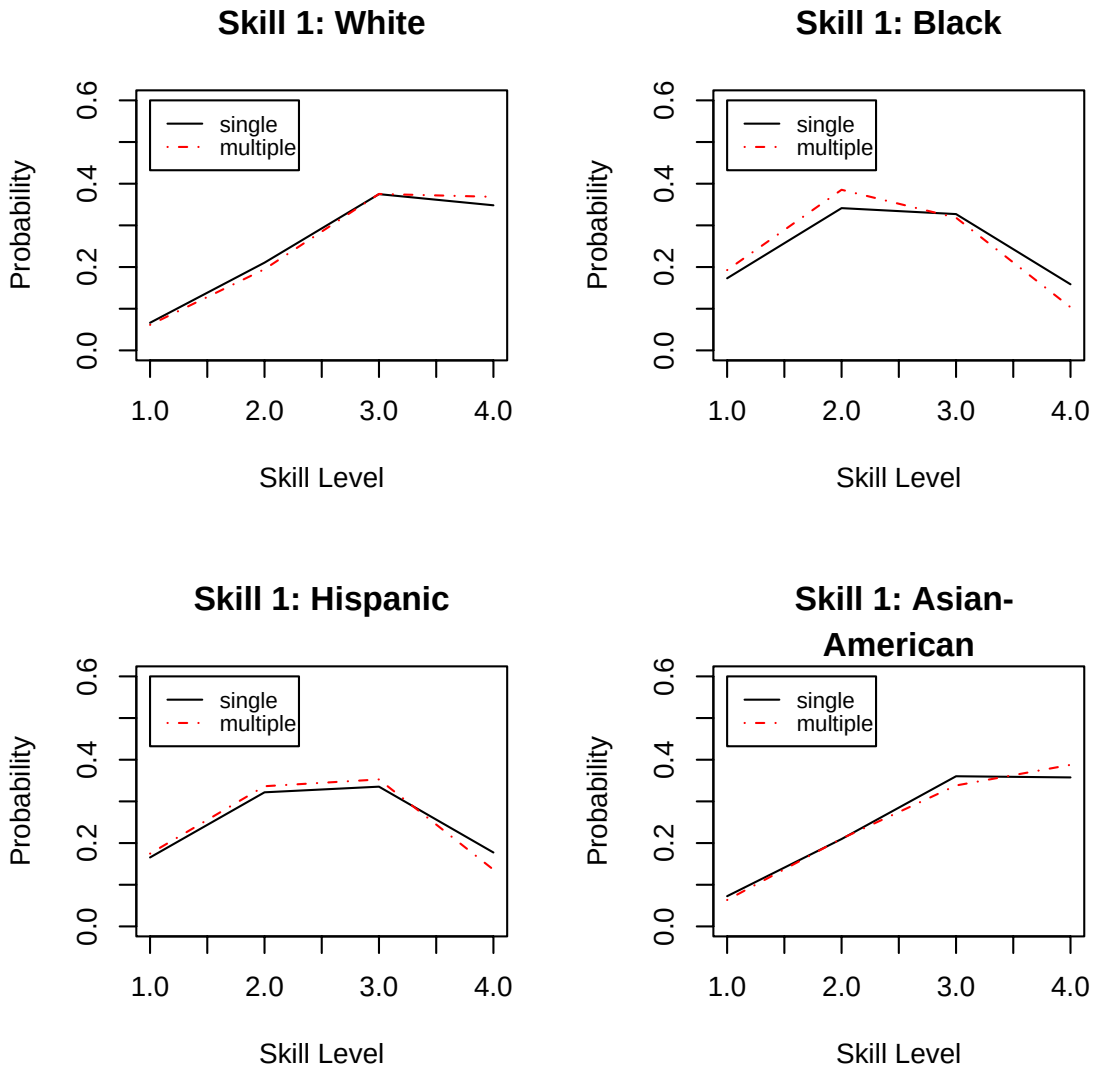

***Figure* 4. Marginal skill distributions: The single-group and race-group analysis of the NAEP 2005 Grade 8 reading data.**

framework. This model can be used with other cognitive diagnosis models, as long as their latent space is discrete.

5.2 Two Assumptions

Both single-group and multiple-group assumptions are used and compared in our analysis. They are comparable to single-ability distribution versus multiple-group IRT models. Each has its

advantages. The results shown in Tables 5 and 4 and Figures 3 and 4 are consistent. There is not much gain using the multiple-group assumption over the single-group assumption when inferences are made on male and female student populations. However, there is a slight difference between these two assumptions when we estimate the skill distributions of racial groups.

The interaction between gender and racial groups was not considered in the analysis. Future research will include study of such interaction effects. In addition, some major reporting subgroups (such as school lunch, parent education, English language learner, etc.) are not included in this analysis, since the skill space would be too large reliable estimation. New models need to be developed in order to reasonably include these conditioning variables in the GDM framework.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19 (6)*, 716-723.

Birch, M. (1963). Maximum likelihood in three-way contingency tables. *Journal of Royal Statistical Society Series B, 26*, 220–233.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

Follmann, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika, 53*, 553–562.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 87*, 476–486.

Haberman, S. (1979). *Analysis of qualitative data*. New York: Academic Press.

Haberman, S. (2005). *Latent class item response models* (ETS Research Rep. No. RR-05-28). Princeton, NJ: ETS.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign, Department of Statistics.

Heinen, T. (1996). *Latent class and discrete latent trait models, similarities and differences*. Thousand Oaks, CA: Sage Publications.

Junker, B., & Sitjsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparameteric item response theory. *Applied Psychological Measurement, 25,* 258–272.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a missing distribution. *Journal of American Statistical Association, 73*, 805–811.

Lang, J. (1996). On the comparison of multinomial and Poisson loglinear model. *Journal of Royal Statistical Society Series B, 58*, 253–266.

Lazarsfeld P. F., & Henry N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*, 523–547.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212.

Palmgren, J. (1981). The Fisher information matrix for loglinear models arguing conditionally in the observed explanatory variables. *Biometrika, 68*, 563–566.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement, 28*, 389–406.

Xu, X., & von Davier, M. (2006). *Applying the general diagnostic model to data from large scale educational surveys* (ETS Research Rep. No. RR-06-08). Princeton, NJ: ETS.